UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

TRABAJO FIN DE GRADO

Técnicas de agrupamiento mediante i-vectors para sistemas automáticos de seguimiento de locutor

David Sánchez Jiménez
Tutor: Javier Franco Pedroso
Ponente: Joaquín González Rodríguez

JUNIO 2018

Técnicas de agrupamiento mediante i-vectors para sistemas automáticos de seguimiento de locutor

AUTOR: David Sánchez Jiménez
TUTOR: Javier Franco Pedroso

Audio, Data Intelligence and Speech
Dpto. de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2018

Resumen (castellano)

En este Trabajo Fin de Grado se propone un sistema de diarización de locutor, cuyo objetivo principal es determinar los instantes de tiempo en que habla cada una de las personas, sin conocimiento a priori de sus identidades, dada una grabación de voz en la que participan varios locutores.

Los sistemas de diarización de locutor dividen el problema en dos subtareas que son la segmentación y el agrupamiento. En primer lugar, se determina los puntos de cambio entre locutores, mediante un detector de actividad que da lugar a segmentos sin identidad asociada. En segundo lugar, se agrupa los segmentos por identidad (aunque en realidad no se conocen las identidades, se asocian los segmentos más parecidos, en base a algún criterio, con una identidad común aunque desconocida).

Mediante el trabajo desarrollado se pretende analizar las distintas técnicas implicadas en el proceso, para dar así con la mejor configuración del sistema que permita obtener la menor tasa de error de diarización. En primer lugar, se estudian y evalúan varias técnicas de scoring como medida de distancia para la etapa de agrupamiento de un sistema de diarización, aislándola de las etapas anteriores. Es en este paso donde hacemos uso de los i-vectors, que representan de forma conjunta la variación debida al locutor y al canal en un mismo espacio de dimensión reducida.

De esta forma, son tres los procedimientos analizados en la etapa de scoring tras el agrupamiento de i-vectors: la distancia coseno, es una medida que se basa en la evaluación del valor del coseno del ángulo comprendido entre ambos vectores; LDA, que es una técnica que reduce la dimensionalidad del espacio de los i-vectors; y PLDA, un modelo generativo para el modelado de i-vectors visto como una versión probabilística de la técnica LDA.

Una vez realizado este análisis, se prueba junto con el resto de etapas del sistema para evaluarlo en condiciones reales, de manera que, empleando una configuración ya optimizada del estudio anterior, la tasa de error de diarización sea la mínima.

En el desarrollo de este trabajo se incluyen pruebas de eficacia y rendimiento de las diversas técnicas sobre la base de datos de audio así como una comparativa entre ellas que nos proporcionará los suficientes datos en el contexto de evaluación.

Palabras clave

Diarización, locutor, voz, audio, segmentación, agrupamiento, detector, actividad, i-vector, distancia, coseno, LDA, PLDA.

Abstract (English)

In this Bachelor Thesis, a speaker diarization system is proposed, whose main objective is to determine the instants of time in which each of the persons speaks, without prior knowledge of their identities, for a recording voice in which several speakers participate.

The speaker diarization systems divide the problem into two subtask such as segmentation and clustering. First, the change points between speakers are determined by means of an activity detector that gives segments without associated identity. Secondly, the segments are grouped by identity (although actually the identities are not know, the most similar segments are associated based on some criterion, with a common identity although unknown).

By means of the developed work, the intention is to analyze the different techniques involved in the process to give the best configuration of the system that allows obtaining the lowest error rate of diarization. First, several scoring techniques are studied and evaluated as a measure of distance for the clustering stage of a diarization system, isolating it from the previous stages. It's in this step where we make use of the i-vectors, which represent of joint form the variation due to the speaker and the channel in the same space of reduced dimension.

In this way, there are three procedures analyzed in the scoring stage after the clustering of the i-vectors: the cosine distance, a measure that is based on the evaluation of the cosine value of the angle between both vectors; LDA, which is a technique that reduces the dimensionality of the space of the i-vectors; and PLDA, a generative model for the modeling of i-vectors seen as a probabilistic version of the LDA technique.

Once this analysis is realized, it is tested together with the rest of the stages of the system to evaluate it in real conditions, so that using a configuration optimized from the previous study, the diarization error rate will be the minimum.

In the development of this Bachelor Thesis there are included test of efficiency and performance of the various techniques based on the audio database as well as the comparative between them that will provide the sufficient information to us in the context of evaluation.

Keywords

Diarization, speaker, voice, audio, segmentation, clustering, detector, activity, i-vector, distance, cosine, LDA, PLDA.

Agradecimientos

A mi tutor, Javier Franco, por darme la oportunidad de realizar este proyecto y guiarme en todo momento en el desarrollo del mismo, por toda la paciencia que ha tenido y estar siempre pendiente. También agradecer a todos los miembros del laboratorio ATVS por su amabilidad.

A mi familia, especialmente a mis padres y a mi hermana, por apoyarme no sólo durante estos cinco años, que se hacen menos duros a vuestro lado, sino en cada aspecto y etapa de mi vida hasta llegar aquí. Sois esenciales en mi vida.

Por último, a mis amigos de toda la vida, que me han sacado una sonrisa en los momentos más difíciles y me han ayudado siempre que lo he necesitado. Sinceramente, no puede haber mejores amigos que vosotros. Y, a las grandes amistades que me llevo de esta carrera, que sin ellos hubiera sido todo muy diferente.

INDICE DE CONTENIDOS

1 Introducción 1
1.1 Motivación 1
1.2 Objetivos 2
1.3 Organización de la memoria 2

2 Estado del arte 5
2.1 Detección de actividad 5
2.1.1 Detección de actividad basada en la energía 5
2.1.2 Detección de actividad basada en modelos 6
2.1.3 Detección de actividad basada en la trayectoria de armónicos 6
2.2 Extracción de características (MFCC) 6
2.3 Segmentación 7
2.3.1 Criterio de Información Bayesiana 7
2.3.2 Razón de probabilidad generalizada 8
2.4 Extracción de los vectores de identidad (i-vectors) 8
2.5 Agrupamiento 10
2.6 Técnicas de normalización de i-vectors 10
2.6.1 Whitening y L-norm 11
2.6.2 Compensación WCCN 11
2.7 Técnicas de scoring en sistemas de diariazión de locutor basados en i-vectors 12
2.7.1 Cosine Distance Scoring (CDS) 12
2.7.2 Linear Discriminant Analysis (LDA) 13
2.7.3 Probabilistic Linear Discriminant Analysis (PLDA) 14

3 Diseño 15
3.1 Introducción 15
3.2 Descripción del sistema de seguimiento de locutor 15
3.2.1 Detector de actividad de voz (VAD) 15
3.2.2 Extracción de características utilizando i-vectors 16
3.2.3 Segmentación 17
3.2.4 Agrupamiento sobre i-vectors 18

4 Desarrollo 19
4.1 Entorno experimental 19
4.1.1 Equipo y programas informáticos 19
4.1.2 Base de datos 20
4.1.2.1 Entrenamiento 20
4.1.2.2 Desarrollo y evaluación 21
4.1.3 Métrica de evaluación 22
4.2 Evolución y particularidades de los experimentos 24
4.2.1 Análisis de parámetros y scoring para agrupamiento de los i-vectors 24
4.2.1.1 Otras casuísticas para LDA 25
4.2.1.2 Otras casuísticas para PLDA 25
4.2.2 Análisis de parámetros para la segmentación 26

5 Integración, pruebas y resultados 29
5.1 Evaluación de las técnicas de scoring..... 29
5.1.1 CDS..... 29
5.1.2 LDA..... 30
5.1.3 PLDA 31
5.2 Evaluación del sistema de diarización de locutor completo configurando la segmentación 33
5.2.1 CDS..... 33
5.2.2 LDA..... 34
5.2.3 PLDA 35

6 Conclusiones y trabajo futuro 37
6.1 Conclusiones..... 37
6.2 Trabajo futuro 37

Referencias 39

INDICE DE FIGURAS

FIGURA 1-1: EJEMPLO DE IDENTIFICACIÓN DE LOS HABLANTES EN SISTEMA DE DIARIZACIÓN 1

FIGURA 2-1: ESPECTROGRAMAS PARA VOZ (A), MÚSICA (B) Y RUIDO (C) 6

FIGURA 2-2: DIAGRAMA DE BLOQUES DEL PROCESO DE EXTRACCIÓN DE LOS MFCC 7

FIGURA 2-3: CONCEPTO DE SUPERVECTOR GMM 9

FIGURA 2-4: ALGORITMO *BOTTOM-UP* 10

FIGURA 2-5: CASOS POSIBLES DE SIMILITUD SEGÚN LA TÉCNICA CDS 12

FIGURA 2-6: AGRUPAMIENTO DE CLASES POR MEDIO DE LDA 13

FIGURA 3-1: ESQUEMA TÍPICO DE UN SISTEMA DE DIARIZACIÓN DE LOCUTOR. 15

FIGURA 3-2: AGRUPAMIENTO BASADO EN LA MEDIDA DE DISTANCIA COSENO. 18

FIGURA 4-1: DISTRIBUCIÓN DE LA BASE DE DATOS DE 3/24 TV SEGÚN SU TIPO DE AUDIO 20

FIGURA 4-2: DISTRIBUCIÓN DE LA BASE DE DATOS DE CARTV SEGÚN SU TIPO DE AUDIO 21

FIGURA 4-3: CÁLCULO DE LA MEDIDA DIARIZATION ERROR RATE A PARTIR DE SUS COMPONENTES 22

FIGURA 4-4: SALIDA EN EL TERMINAL DE LA EVALUACIÓN DEL SISTEMA 23

FIGURA 4-5: REPRESENTACIÓN DE LA SEGUNDA ETAPA DE ESTUDIO DEL SISTEMA 27

INDICE DE TABLAS

TABLA 4-1: DISTRIBUCIÓN CUANTITATIVA DE LA BASE DE DATOS 21

TABLA 5-1: TASA DE ERROR DE DIARIZACIÓN DER EN FUNCIÓN DE LA NORMALIZACIÓN PARA LOS DATOS DE DESARROLLO 29

TABLA 5-2: TASA DE ERROR DE DIARIZACIÓN DER PARA VALORES DE CUTOFF ADYACENTES A LOS MÍNIMOS ENCONTRADOS PARA LOS DATOS DE DESARROLLO 29

TABLA 5-3: TASA DE ERROR DE DIARIZACIÓN DER PARA LA MEJOR CONFIGURACIÓN CON LOS DATOS DE EVALUACIÓN (CDS) 30

TABLA 5-4: TASA DE ERROR DE DIARIZACIÓN DER CON DISTINTO NÚMERO DE DIMENSIONES DEL ESPACIO PROYECTADO PARA LOS DATOS DE DESARROLLO (LDA) 30

TABLA 5-5: TASA DE ERROR DE DIARIZACIÓN DER PARA LA MEJOR CONFIGURACIÓN CON LOS DATOS DE EVALUACIÓN (LDA) 31

TABLA 5-6: TASA DE ERROR DE DIARIZACIÓN DER CON DISTINTOS TAMAÑOS DE MATRICES PARA LOS DATOS DE DESARROLLO (PLDA) 32

TABLA 5-7: TASA DE ERROR DE DIARIZACIÓN DER CON TAMAÑO DE MATRIZ ADAPTATIVO PARA LOS DATOS DE DESARROLLO 32

TABLA 5-8: TASA DE ERROR DE DIARIZACIÓN DER PARA LA MEJOR CONFIGURACIÓN CON LOS DATOS DE EVALUACIÓN (PLDA) 33

TABLA 5-9: TASA DE ERROR DE DIARIZACIÓN DER CONFIGURANDO LA SEGMENTACIÓN CON LOS DATOS DE DESARROLLO (CDS) 33

TABLA 5-10: TASA DE ERROR DE DIARIZACIÓN DER SOBRE EL SISTEMA DE DIARIZACIÓN COMPLETO CON LOS DATOS DE EVALUACIÓN (CDS) 34

TABLA 5-11: TASA DE ERROR DE DIARIZACIÓN DER CONFIGURANDO LA SEGMENTACIÓN CON LOS DATOS DE DESARROLLO (LDA) 34

TABLA 5-12: TASA DE ERROR DE DIARIZACIÓN DER SOBRE EL SISTEMA DE DIARIZACIÓN COMPLETO CON LOS DATOS DE EVALUACIÓN (LDA) 35

TABLA 5-13: TASA DE ERROR DE DIARIZACIÓN DER CONFIGURANDO LA SEGMENTACIÓN CON LOS DATOS DE DESARROLLO (PLDA) 35

TABLA 5-14: TASA DE ERROR DE DIARIZACIÓN DER SOBRE EL SISTEMA DE DIARIZACIÓN COMPLETO CON LOS DATOS DE EVALUACIÓN (PLDA) 35

1 Introducción

1.1 Motivación

El habla sigue siendo una de las formas más utilizadas por los seres humanos para comunicar las ideas y transmitir información al mundo. De hecho, la cantidad de información disponible por medio del habla (a través de diferentes plataformas como teléfono, televisión, radio, reuniones, etc.) que se está almacenando es considerablemente elevada y, precisamente este gran número de datos junto con la importancia vital que tiene la voz en nuestra sociedad, hace necesario el desarrollo de sistemas que analicen la voz.

La diarización de locutores ha surgido como un tema cada vez más importante en la gestión de contenidos. Mientras que el reconocimiento de locutor implica el reconocimiento de la identidad de una persona para saber qué se dice, la diarización trata de resolver "quién" habló y cuándo habló. Más formalmente, consiste en la división de un fichero de audio de entrada compuesto por varios locutores en segmentos uniformes y posteriormente agrupar las partes resultantes en base a la identidad de cada locutor. Como se muestra en la **Figura 1-1**, esto es de gran utilidad puesto que podemos etiquetar por segmentos qué locutor interviene en cada momento o intervalo de tiempo.

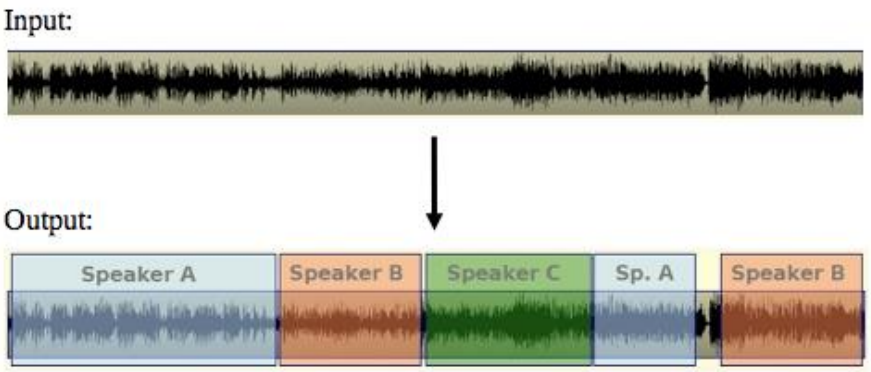

Figura 1-1: Ejemplo de identificación de los hablantes en sistema de diarización. | Fuente: [8]

Saber cuándo habla un locutor en una grabación de audio por tanto es muy útil en sí mismo, pero también, esta etapa de diarización es un paso de procesamiento importante en muchas tareas. Por ejemplo, su combinación con el reconocimiento automático de locutor facilita información adicional para las transcripciones ya que atribuye regiones de hablante a un archivo de audio y permite agrupar los segmentos de voz de cada uno de los hablantes para proporcionar la verdadera identidad del locutor en el flujo de audio. También cabe destacar la posibilidad de aislar los tramos en que habla un único locutor para poder adaptar un sistema de reconocimiento de voz a ese locutor.

Dada las múltiples aplicaciones prácticas de los sistemas de diarización de locutor como "navegar" por una grabación de gran tamaño, ir a los tramos donde habla alguno de los locutores, saltar de uno a otro, etc. resulta de gran importancia realizar un estudio en este trabajo fin de grado sobre las diferentes técnicas utilizadas para obtener la mejor respuesta del sistema.

1.2 Objetivos

El objetivo principal de este TFG es el de examinar el sistema de diarización de locutor, dada la base de datos empleada en la evaluación Albayzín, para su posterior mejora en cuanto a rendimiento. Esta evaluación consiste en la creación de etiquetas donde se indican los intervalos en los que participan diferentes locutores para grabaciones de audio que forman parte de datos de entrenamiento y desarrollo del sistema.

Para ello, primero vamos a comprender el funcionamiento y comparar tres técnicas en el estado del arte para realizar el proceso de agrupamiento como son cosine distance, LDA[1] (técnica de reducción de dimensionalidad) y PLDA[2], junto con procedimientos de normalización, para ajustar los parámetros del modelo respectivo de forma que se obtenga la menor tasa de error posible.

Una vez realizado, el segundo paso es observar el funcionamiento completo de la diarización de locutor. Esto conlleva, estudiar también la segmentación y la detección de actividad para ver cómo funciona el sistema en una situación real.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

1. **Capítulo 1: Introducción.**

   Este primer capítulo detalla brevemente la motivación para el desarrollo de este proyecto, así como, los objetivos marcados durante la ejecución del TFG. Además, se realiza también una estructura acerca de los contenidos presentes en la memoria.

2. **Capítulo 2: Estado del arte.**

   En este capítulo se presenta el estado del arte actual de las técnicas de agrupamiento en los sistemas de diarización de locutor. Primero, se describen las etapas correspondientes a estos sistemas. Por último, se analizan las diferentes técnicas empleadas de scoring a partir de los i-vectors y normalizaciones.

3. **Capítulo 3: Diseño.**

   Este capítulo incluye una descripción de las características del tipo de sistema empleado en la ejecución de los experimentos, particularizando en las correspondientes a cada etapa.

1 Linear Discriminant Analysis
2 Probabilistic Linear Discriminant Analysis

4. **Capítulo 4: Desarrollo.**

   Esta sección, detalla la base de datos empleada durante el proyecto. Asimismo, describe la metodología del proceso realizado hasta la consecución de los resultados además de una documentación de las herramientas utilizadas.

5. **Capítulo 5: Integración, pruebas y resultados.**

   En este capítulo se presentan los resultados obtenidos durante la evaluación de las técnicas analizadas para poder así realizar una comparación entre ellas. De igual forma, se explica los experimentos realizados para dar con la mejor configuración de los parámetros.

6. **Capítulo 6: Conclusión y trabajo futuro.**

   Por último, se hace una valoración final con las conclusiones extraídas del proyecto realizado y se detallas futuras líneas a seguir para poder mejorar nuestro sistema.

2 Estado del arte

Los sistemas automáticos de seguimiento de locutor facilitan reconocer los periodos en que participan distintos hablantes en una grabación de audio. Es por ello que su uso es básico para aplicaciones como indexación de contenidos o simplemente como etapa de preprocesado para aquellas que necesiten a su entrada la intervención de un único locutor, como es el caso de reconocimiento de locutor o reconocimiento de voz adaptada al locutor. Todas ellas hacen uso de los i-vectors, ya que son el fundamento de los sistemas en el estado del arte en muchas de estas aplicaciones.

En este capítulo se presenta el estado del arte en los sistemas de diarización de locutor. Estos mismos se basan en tres bloques principales: detección de actividad de voz (VAD)[3], segmentación y agrupamiento. Se hará especial hincapié a las técnicas empleadas en la etapa de agrupamiento analizando las características de cada una de ellas.

2.1 Detección de actividad

Un aspecto transcendental en los sistemas de diarización de locutor es la detección de actividad de voz (VAD). Un preciso etiquetado del habla es de vital importancia para el rendimiento del sistema de manera que se determine correctamente los segmentos con voz y sin voz. No obstante, hay que tener en cuenta, que los segmentos sin voz, además de poder tratarse de segmentos de silencio, pueden ser ruido de fondo, música, o solapamientos entre ellos que afecta al proceso y hace que el desarrollo de este módulo sea determinante.

La mayoría de dificultades vienen al etiquetar cuando no hay voz y no etiquetar cuando en realidad si hay presencia de habla. Estos errores, junto con la importancia de optimizar la detección de actividad de voz, dan lugar a numerosas estrategias que siguen diferentes recorridos.

2.1.1 Detección de actividad basada en la energía

La estrategia más sencilla de implementar es la que se fundamenta en el cálculo de la energía de la señal. Para ello, se fija un umbral de energía, que indica zona de habla cuando la magnitud de la amplitud de la señal supere el umbral y zona de silencio en caso contrario.

Los resultados que se obtienen siguiendo esta técnica no son efectivos puesto que la música o ruido anteriormente mencionados pueden superar la energía de ese umbral lo que da lugar a etiquetas incorrectas, al clasificar como voz segmentos que en realidad son otros eventos acústicos.

3 Voice Activity Detection

2.1.2 Detección de actividad basada en modelos

Una estrategia más efectiva y a su vez compleja, consiste en la utilización de modelos que caractericen todos los posibles tipos de audio. En este caso, se utilizan modelos de mezclas de gaussianas (GMMs[4]) que contemplen toda la casuística posible. Con esto se consigue calcular la verosimilitud de cada trama de audio para cada uno de los casos, y se elige el que mayor probabilidad tenga. Un ejemplo de este procedimiento es el sistema LIUM, que presenta ocho categorías distintas para modelar todas las clases [16].

2.1.3 Detección de actividad basada en la trayectoria de armónicos

El procedimiento de detección de actividad basado en la trayectoria de los armónicos está fundamentado en el reconocimiento de voz y música basado en características espectrales. Esta representación presenta buenos resultados ya que la señal de voz tiene propiedades únicas en la posición de algunos armónicos que la diferencian de la música y el ruido.

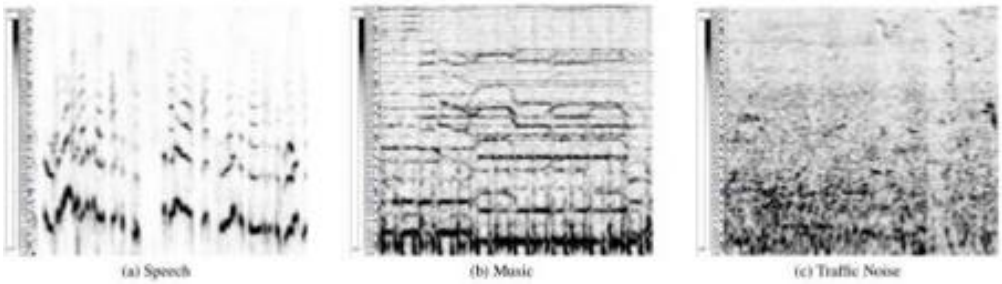

Figura 2-1: Espectrogramas para voz (a), música (b) y ruido (c). | Fuente: [18]

A diferencia de la señal de voz como vemos en la **Figura 2-1**, la música se caracteriza por armónicos prolongados en el tiempo en la zona baja del espectro, mientras que el ruido tiene un patrón aleatorio.

2.2 Extracción de características (MFCC[5])

En nuestro trabajo se ha empleado los coeficientes MFCC para la extracción de las características frecuenciales de los segmentos. Estos representan la información propia de cada hablante de forma compacta. Los pasos para su obtención son los indicados a continuación:

1. En primer lugar, se divide el audio en segmentos de 20 ms y normalmente, con 10 ms de solapamiento entre ellos. Cada uno de ellos se pasa por una ventana tipo Hamming para impedir efectos no deseados en los bordes.

4 Gaussian Mixture Models
5 Mel Frecuency Cepstral Coefficients

2. Se realiza la FFT[6] sobre cada segmento para extraer sus frecuencias y se aplica un banco de filtros Mel.

3. Se calcula su logaritmo y se realiza la transformada del coseno obteniendo un vector de coeficientes por cada segmento.

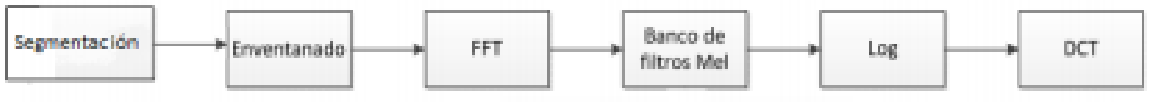

Figura 2-2: Diagrama de bloques del proceso de extracción de los MFCC

Como resultado se obtiene para cada locución su representación con un número variable de vectores de características MFCC.

2.3 Segmentación

La segmentación es el proceso encargado de determinar los límites de los segmentos homogéneos etiquetados como voz procedentes del módulo anterior. En nuestro trabajo, el objetivo es el identificar como fronteras los cambios de locutor entre segmentos consecutivos. Al no emplear ningún tipo de información previa acerca la identidad o el número de locutores presentes en el fichero de audio, el propósito es localizar los cambios de locutor.

Para realizar la detección de cambio de locutor, generalmente se emplea una medida de distancia acústica que evalúa la similitud entre dos ventanas adyacentes que se desplazan recorriendo todo el audio, identificando los puntos donde se produce un cambio o límites entre segmentos a partir de un umbral fijado.

2.3.1 Criterio de Información Bayesiana

El Criterio de Información Bayesiana o BIC[7] es la técnica más empleada en segmentación de locutores. Es utilizado para seleccionar el modelo más apropiado en cada caso a partir de un criterio probabilístico penalizado por la complejidad del modelo. Por ello, se utiliza para decidir si los segmentos analizados pertenecen a un solo locutor o, por el contrario, se ha detectado cambio de locutor.

Para determinar si hay un cambio de locutor se evalúa la hipótesis asumiendo que los datos de los segmentos quedan representados por un único modelo frente a la hipótesis que supone que se adecuan mejor con dos modelos, de la siguiente manera:

$$\Delta BIC = BIC_{H2} - BIC_{H1} = (ni + nj)\ln|\Sigma| - ni\ln|\Sigma i| - nj\ln|\Sigma j| - \lambda P \quad (2.1)$$

6 Fast Fourier Transform
7 Bayesian Information Criterion

donde Σ es la matriz de covarianza, ni y nj el número de vectores de características en los respectivos conjuntos *ci* y *cj*, *P* es la penalización y λ es el umbral de decisión.

2.3.2 Razón de probabilidad generalizada

La Razón de Probabilidad Generalizada o GLR[8] calcula la relación de verosimilitudes entre la hipótesis que afirma que el conjunto de características proviene del mismo locutor frente a la hipótesis que indica que procede de diferente locutor. Esta relación de verosimilitudes entre las hipótesis viene dada por:

$$GLR = \frac{H_1}{H_2} = \frac{P(X,M)}{P(X_i,M_i)P(X_j,M_j)} \quad (2.2)$$

donde X es el conjunto de datos de los segmentos y, M_i y M_j son los modelos que se ajustan mejor a los datos.

2.4 Extracción de los vectores de identidad (i-vectors)

Los i-vectors son la base en el estado del arte de los sistemas de seguimiento de locutor debido a una representación vectorial más compacta de la gran cantidad de las características acústicas dado un fichero de voz de entrada de duración variable. Su desarrollo conllevó grandes avances en el reconocimiento de locutor [1] con grandes ventajas respecto a la técnica JFA[9].

La técnica JFA pretende modelar la variabilidad del sistema mediante la separación de la voz en dos subespacios: el espacio de locutor (*s*) y el de canal (*c*). El resultado es un supervector *M*, compuesto por la suma de ambos espacios, tal como refleja la fórmula:

$$M = s + c \quad (2.3)$$

Por un lado, el espacio *c* se define como

$$c = Ux \quad (2.4)$$

Donde *U* es la matriz *eigenchannel* y *x* es el factor dependiente del canal. Por otro lado, el espacio *s* se descompone en

$$s = m + Vy + Dz \quad (2.5)$$

Donde *m* es un supervector independiente de locutor y canal formado por la concatenación de las medias de las mezclas de un Universal Background Model (UBM), *V* es la matriz *eigenvoice*, *D* es una matriz diagonal que obtiene la información residual e *y* y *z* son vectores dependientes del locutor en su correspondiente subespacio.

8 Generalized Likelihood Ratio
9 Join Factor Analysis

Por el contrario, la técnica de los i-vectors se fundamenta en definir un único espacio que comprenda conjuntamente información de locutor y canal, llamado espacio de variabilidad total que contiene variabilidad de locutor y del canal de manera simultánea. Se descompone de la siguiente forma:

$$M = m + Tw \quad (2.6)$$

Aquí, *m* es un supervector independiente del locutor y del canal extraído del UBM, *T* es una matriz rectangular de bajo rango y *w* es el i-vector que proyecta los datos en el espacio de variabilidad total.

Para obtener el i-vector que modele cada uno de los segmentos del fichero de audio, primeramente deberán extraerse las características frecuenciales de los segmentos siguiendo el procedimiento visto en el Apartado 2.2.

A continuación, se encuentra la etapa de entrenamiento que consiste en elaborar un modelo del locutor a partir de los vectores de características MFCC. Para ello, se ha empleado el modelado mediante GMM, que es un modelo probabilístico que asume que los datos son generados por una mezcla de un número finito de distribuciones gaussianas de parámetros desconocidos, junto con el modelo universal UBM que modela las características comunes a todos los locutores.

Tras el entrenamiento GMM-UBM, se extraen estadísticas a partir de las características de la entrada, y se entrena el subespacio de variabilidad total, que se puede entender como una matriz de proyección de los supervectores en i-vectors, consiguiendo así almacenar la información en un espacio reducido, en nuestro caso, de dimensión 600.

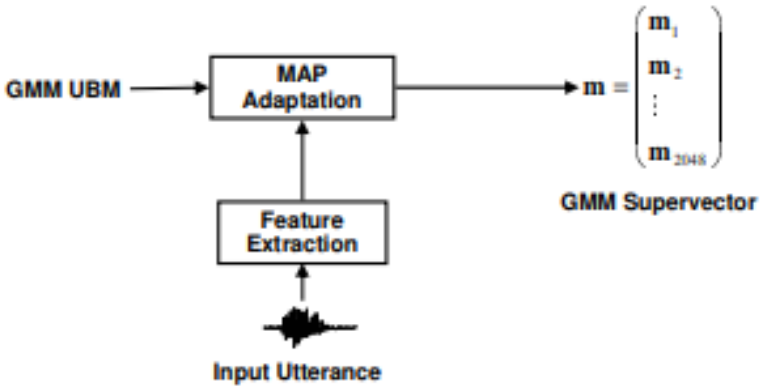

Figura 2-3: Concepto de supervector GMM. | Fuente: [21]

Por último, se extraerán a partir de los segmentos de entrenamiento de entrada y evaluación, los i-vectors (de dimensión 600 en nuestro desarrollo) que los caractericen. Una vez obtenidos los i-vectors deberán compararse con los segmentos de análisis para obtener una puntuación o score según su similitud con este.

2.5 Agrupamiento

El cometido del bloque de agrupamiento es identificar los segmentos que corresponden a un mismo locutor y los que se identifican a diferentes locutores.

En nuestro sistema se ha hecho uso del Agrupamiento Jerárquico Aglomerativo (AHC[10]) o estrategia *bottom-up* [17]. Este esquema se basa en la unión iterativa de los i-vectors que den lugar a un score más alto al compararlos (independientemente de la técnica de scoring empleada). Cuando se realiza la unión, se considera que el cluster resultante se representa mediante la media y continua hasta que se cumple un criterio de parada, en nuestro caso, *cutoff*.

Para ello, inicialmente se supone una sobre-segmentación de la señal de audio en un número mayor de segmentos al número real de hablantes. En primer lugar, considerando cada segmento como cluster único y posteriormente, agrupando en función de la distancia hasta llegar al *cutoff*.

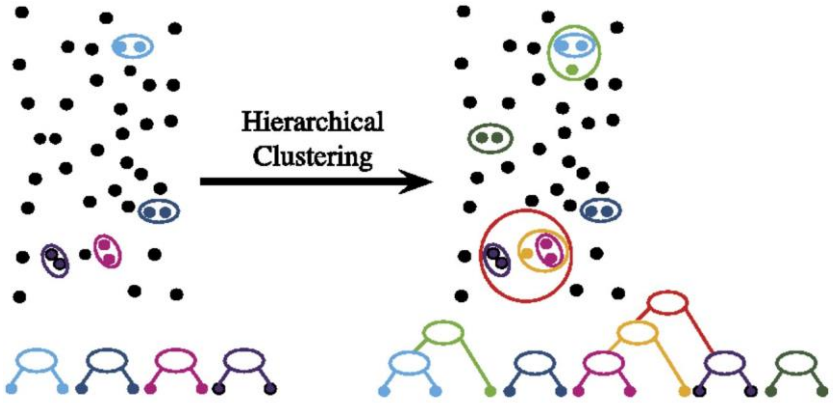

Figura 2-4: Algoritmo *bottom-up.* | Fuente: [22]

2.6 Técnicas de normalización de i-vectors

Una vez extraídos los i-vectors, el objetivo será obtener una medida de similitud entre los segmentos modelados por los i-vectors y los segmentos de análisis o referencia. Existen varias formas de obtener una puntuación o score, pero en este trabajo se presentan: distancia coseno y PLDA.

LDA es una técnica también utilizada en nuestro sistema pero por sí sola no proporciona una score. Es más bien una técnica de "compensación de variabilidad". Por ejemplo, PLDA emplea LDA para modelar la variabilidad en un espacio de dimensionalidad reducida.

Todos ellos usan herramientas de normalización que conviene explicar antes de focalizar y detallar cada una de las técnicas de scoring de forma extensa.

10 Agglomerative Hierarchical Clustering

2.6.1 Whitening y L-norm

El uso de whitening está bastante estandarizado en reconocimiento de locutor. Sin embargo, en diarización de locutor no es tan frecuente. Su uso junto con la normalización de longitud mejora el rendimiento de las técnicas de scoring.

Se consideró necesaria la normalización de longitud de los i-vectors para hacer que el clustering funcionara con éxito. La normalización consiste en dividir cada i-vector por su módulo, así que el vector resultante es de longitud unitaria. Este paso fue muy útil para eliminar los sesgos entre los conjuntos de datos de entrenamiento, desarrollo y prueba [2].

Previo a la normalización, los i-vectors deben estar centrados y 'blanqueados'. La operación de centrado consiste en la resta de la media global de todos los i-vectors entrenados a cada i-vector, por lo que terminan centrados en el origen de las coordenadas. La técnica de whitening es una transformación que normaliza a varianza unidad. Después de estas dos operaciones, los i-vectors no están correlacionados.

Es después de realizar l-norm cuando se colocan en la hiperesfera unidad distribuidos uniformemente alrededor del origen. Si no se aplicara el centrado y whitening, los i-vectors se concentrarían en una pequeña región de la hiperesfera unidad, sin ningún poder discriminatorio para la clasificación. Es por ello la importancia de emplear ambas técnicas juntas, que como veremos en el Capítulo 5, conseguiremos así los mejores resultados.

2.6.2 Compensación WCCN

La normalización de la covarianza intra-clases o WCCN[11] es una técnica que se emplea para la compensación de la variabilidad de sesión en el espacio i-vector. Esto ayuda a reducir las variaciones provocadas en un mismo locutor, a partir de la covarianza promedio de unos hablantes, definida de la siguiente manera:

$$\Phi(x) = B'x, \quad (2.7)$$

donde x es un i-vector y B es una matriz de proyección obtenida de la descomposición:

$$W^{-1} = BB' \quad (2.8)$$

En este caso, W es la covarianza promedio de la variabilidad intra-clase de un conjunto de locutores y sigue la siguiente fórmula [3]:

$$W = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{n_l}\sum_{i=1}^{n_l}(x_i^l - x_l)(x_i^l - x_l)' \quad (2.9)$$

Donde L es el número de locutores totales, n_i es el total de segmentos donde hay voz por cada hablante, x_i es el vector que resulta del cálculo de la media de todos los i-vectors correspondientes a un solo locutor y x_i^l es el i-ésimo i-vector del locutor.

11 Within Class Covariance Normalization

En nuestro sistema, esta técnica se ha usado junto con la técnica LDA. Esto es posible porque los i-vectors utilizados para estimar la matriz de proyección dada por WCCN están compensados en función de la variabilidad de sesión con la matriz de proyección de LDA.

2.7 Técnicas de scoring en sistemas de diarización de locutor basados en i-vectors

Tras obtener los i-vectors, existen varias maneras para lograr un score que mida la similitud entre los obtenidos por extracción y los de referencia contra los que debemos comparar. Se van a analizar las técnicas más empleadas en los sistemas de diarización de locutor.

2.7.1 Cosine Distance Scoring (CDS)

La obtención de la puntuación en este método resulta a partir del ángulo que forman dos i-vectors no nulos. Cuanto más pequeño sea el ángulo que separa dichos vectores más parecidos serán debido a su proximidad, lo que conlleva a mayor probabilidad de la hipótesis de que pertenezcan a un mismo locutor. La score se define de la siguiente manera:

$$score\ (iv1,iv2) = \frac{iv1 * iv2}{\|iv1\| * \|iv2\|} \quad (2.10)$$

donde *iv1* e *iv2* son los i-vectors en cuestión. Se basa en que existe información no relevante con el locutor, como pueda ser el canal o la sesión, en la magnitud de los i-vectors, pero realmente es el ángulo el que distingue a las personas. Esto hace que el método sea robusto, poco complejo y rápido computacionalmente.

Como el coseno puede oscilar entre el intervalo [-1,1], según el valor que alcance indicará:

- Si es cercano a 1, los i-vectors están en la misma dirección, por lo que la probabilidad de que se trate del mismo hablante o locutor es alta.
- Si es cercano a -1, los i-vectors son resultados opuestos, lo que hace indicar que se trata de locutores diferentes.
- Si es similar a 0, los i-vectors son ortogonales entre ellos. En este caso intermedio indicará que hay una cierta similitud.

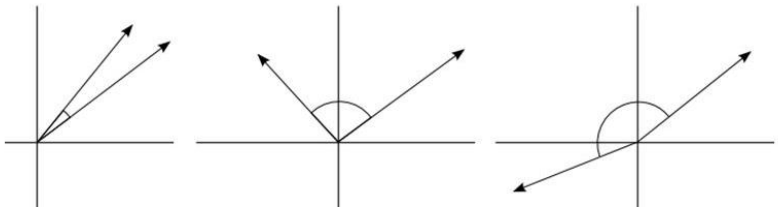

Figura 2-5: Casos posibles de similitud según la técnica CDS

Como hemos comentado anteriormente, esto permite que la forma de implementar el algoritmo sea bastante rápida y sencilla.

2.7.2 Linear Discriminant Analysis (LDA)

El análisis discriminante lineal es una generalización del discriminante lineal de Fisher, cuya idea básica es extraer combinaciones lineales de características, donde las medias de las clases estén lejos entre ellas y la varianza dentro de cada clase sea pequeña [4].

La puntuación en esta técnica se calcula de manera análoga a la vista en el apartado anterior, esto es, mediante el cálculo de la distancia coseno. A diferencia de CDS, en este caso, los i-vectors se proyectan a un espacio de menor dimensionalidad a través de la matriz de proyección *A* empleada en LDA.

El objetivo de la técnica LDA consiste en la búsqueda de una nueva base ortogonal en el espacio de características (i-vectors) de igual o inferior tamaño dimensional que mejor recoja la discriminación entre locutores. El primer paso es calcular la separabilidad entre los diferentes locutores (clases). El segundo paso es calcular la distancia entre la media y las muestras de cada locutor. El tercer paso es construir dicho espacio dimensional menor que minimiza la variabilidad intra-locutor y maximiza, a su vez, la variabilidad inter-locutor.

Para desplazar los i-vectors al nuevo espacio, LDA hace uso de una matriz de proyección *A* formada por los mejores *eigenvectors* de la siguiente ecuación:

$$M_b v = \lambda * M_w v \quad (2.11)$$

Donde λ es la matriz diagonal de *eigenvalues* y, M_b y M_w son las matrices que se corresponden con la covarianza inter-locutor e intra-locutor. Debido a que LDA hace uso tanto de la intra-variabilidad y la inter-variabilidad de locutor, esta técnica necesita etiquetas de locutor.

Tras obtener nuestro nuevo espacio dimensional y desplazar los i-vectors al mismo, es muy común realizar una compensación de la variabilidad, es decir, es frecuente emplear la técnica WCCN (LDA+WCCN) [5]. Esto es posible porque los i-vectors empleados tras la aplicación de la matriz *A* son compensados gracias a la matriz *B* de WCCN.

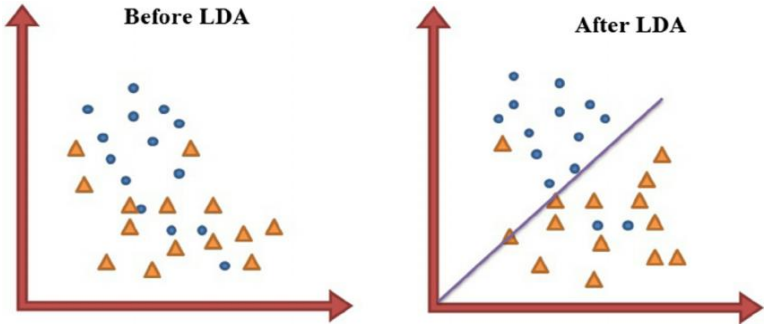

Figura 2-6: Agrupamiento de clases por medio de LDA. | Fuente: [23]

Como vemos en la **Figura 2-6**, con esta técnica buscamos una combinación lineal que nos permita reducir la dimensión del problema de tal manera que nos sea más fácil diferenciar dos o más clases, en nuestro caso, locutores.

2.7.3 Probabilistic Linear Discriminant Analysis (PLDA)

Tanto PLDA como LDA se ocupan en la intra-variabilidad e inter-variabilidad de locutor. Estos métodos estiman las direcciones en el espacio de total variabilidad (TV) que maximiza la discriminación del hablante. Este modelo se puede ver análogo al método de JFA [6], donde se dividen las variaciones debidas al canal, locutor y componentes residuales de la siguiente forma:

$$M = m + Vy + Ux + Dz \quad (2.12)$$

Donde *V, U* y *D* son las matrices dependientes del locutor, canal y componente residual y partir de estas, se consiguen los factores de viabilidad entre locutores *y*, canales *x*, y componentes residuales *z*.

En cambio hay una diferencia entre ellos, y es que mientras JFA está aplicada a supervectores, PLDA lo hace sobre los i-vectors [7]. Dadas *J* locuciones de *I* hablantes con i-vector $x_{ij}: i=1,.....,I;\ j=1,....,J$ estos vectores pueden descomponerse de la siguiente forma:

$$x_{ij} = m + Vh_i + Uw_{ij} + \varepsilon_r \quad (2.13)$$

donde *m* es la media de todos los datos de entrenamiento del modelo, *V* es una matriz que representa el subespacio del locutor, *U* es una matriz que modela el subespacio de la sesión y ε_r es una variable aleatoria con distribución normal que representa el resto de la variabilidad y se le nombra como término residual. Por su parte, h_i y w_{ij} son vectores formados por variables aleatorias que contienen los factores del hablante y canal.

Por ello, dos segmentos de audio serán de un mismo locutor cuanto más similares sean entre sí sus vectores *h*.

En la etapa de entrenamiento, a partir de los i-vectors se estiman los parámetros que definen el modelo PDLA. Esto se realiza a partir del algoritmo de Expectación-Maximización (E-M) [20] que encuentra estimaciones de parámetros de máxima verosimilitud en modelos probabilísticos alternando entre dos pasos, expectativa (E) y maximización (M). E-M hace uso del modelo finito de mezclas gaussianas y estima un conjunto de parámetros iterativamente hasta alcanzar un valor de convergencia deseado. El algoritmo funciona de la siguiente manera:

1. Estima los parámetros iniciales de los modelos: media y desviación típica.

2. Iterativamente mejora los parámetros con los pasos E y M.

- En el paso E, calcula la posibilidad de pertenencia para cada grupo (en nuestro caso, locutor) en función de los valores de los parámetros iniciales.
- En el paso M, vuelve a calcular los parámetros en función de la nueva probabilidad de pertenencia.

3. Asigna a cada locutor al cluster con el que tenga mayor posibilidad de pertenencia.

3 Diseño

3.1 Introducción

El diseño de un sistema de seguimiento de locutor no sigue siempre el mismo esquema (ej.: VAD y extracción de características pueden alternarse o realizarse en paralelo) siendo numerosas las configuraciones que se puedan realizar y las técnicas que se pueden emplear en cada etapa.

En este capítulo, tras analizar las distintas técnicas que se usan para resolver el problema, se concretan las técnicas empleadas en nuestro sistema así como una descripción del diseño del mismo.

3.2 Descripción del sistema de seguimiento de locutor

El sistema empleado de diarización de locutores en este Trabajo Fin de Grado está basado en las siguientes fases: extracción de características, detección de actividad (VAD), segmentación y agrupamiento.

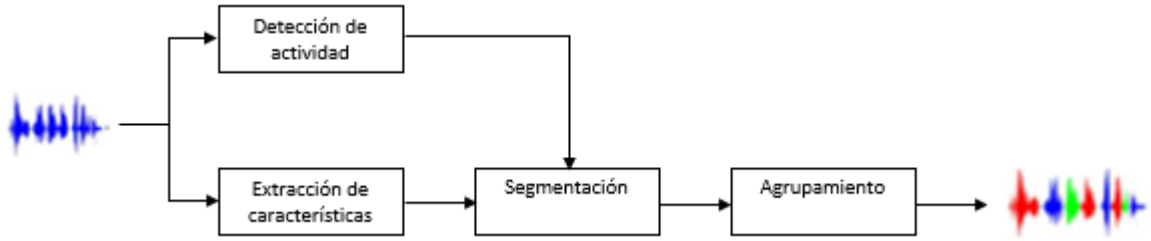

Figura 3-1: Esquema típico de un sistema de diarización de locutor.

Como resultado, indicado en la **Figura 3-1**, obtenemos una señal de audio donde se identifican el principio y fin de cada segmento asociado a un locutor que puede o no participar en el resto de la grabación. En los siguientes apartados se detallará los parámetros escogidos en cada fase del sistema para solucionar el problema.

3.2.1 Detector de actividad de voz (VAD)

La detección de actividad en el sistema de diarización se realiza a partir de un módulo que clasifica los segmentos del audio de entrada como con y sin voz. Esto es de capital importancia para el rendimiento de nuestro sistema y permite excluir los segmentos donde no hay voz en el análisis posterior.

En nuestro caso, el detector de actividad de voz [18] se basa en un análisis espectral del fichero de audio, en concreto, de la posición de los armónicos en un espectrograma.

Con esta representación, se observan características específicas de la señal de voz que lo distinguen de la música o ruido. Para ello, se efectúan las siguientes acciones:

1. Se divide la señal a través de ventanas Hamming de 30 ms cada 10 ms, y se realiza la FFT por cada una.
2. Se divide la energía que hay en cada ventana en un eje de frecuencias distribuido en 6 octavas, cada octava de 40 columnas logarítmicas.

Por tanto, a partir de la trayectoria de los armónicos (en un espectrograma de segmentos de voz, estos armónicos se identifican por estar sostenidos durante periodos considerablemente largos) se puede realizar la detección de voz. Para ello, se calcula la correlación cruzada a partir del espectrograma logarítmico sobre la señal de voz de la siguiente manera:

$$R_{xy}(l) = \sum_l X_t X_{t+offset}(l) \quad (3.1)$$

Donde X_t y $X_{t+offset}$ son tramas de audio consecutivas (separadas por un offset $\in$[-4,4]) mientras que $l \in$ [-10,10] es el desplazamiento frecuencial. Esto se realiza porque para decidir si hay o no voz se define el valor:

$$R = r_{xcorr} - r \quad (3.2)$$

Siendo r_{xcorr} el máximo valor de la correlación y r el valor de la correlación para un desplazamiento frecuencial de valor nulo. Así, podemos distinguir:

- Si la señal es musical, viendo que tiene patrones espectrales horizontales, el valor máximo de correlación para tramas consecutivas será cuando l=0, por lo que, el valor de R=0.
- Si la señal es de voz, al tener trayectoria curva y poco firme en sus armónicos, el valor máximo de correlación para tramas consecutivas será cuando $l \neq 0$ y por tanto, el valor de R > 0.

3.2.2 Extracción de características utilizando i-vectors

Esta etapa se realiza de manera paralela a la detección de actividad. Es de vital importancia puesto que distinguir cada uno de los locutores que participan en un fichero de audio es una tarea complicada que afecta directamente a la tasa de error de diarización.

Para el análisis de voz, el método empleado son los coeficientes MFCC. Se extrae las características a través de un vector de 20 coeficientes MFCC por medio de ventanas Hamming de 20 ms con un solapamiento del 50%. Esto último es útil para asegurar que las discontinuidades que puedan existir en el fichero no se pasen por alto y se analicen en el desplazamiento siguiente.

Cada 10 ms se tiene una ventana donde se calcula un vector MFCC a través de un banco de 25 filtros Mel triangulares equiespaciados.

A partir de estos coeficientes se extraen i-vectors de 600 dimensiones con un extractor entrenado sobre la base de datos proporcionada en la evaluación Albayzín 2010 descrita en el Capítulo 4.

Principalmente este bloque tiene como función representar un conjunto de vectores de características en un único vector que guarda la información discriminativa del locutor de manera eficaz. De esta forma, se extraen i-vectors de tamaño 600 a partir de cada segmento de longitud variable que sale de la etapa de segmentación.

3.2.3 Segmentación

La segmentación es una parte fundamental en los sistemas de seguimiento de locutor. Su objetivo es dividir la parte del audio clasificada como voz, que proviene del módulo de detector de actividad, en segmentos de cada locutor. Para conseguir esto se realiza una comprobación de una hipótesis para la similitud entre segmentos en dos ventanas solapadas adyacentes de 5 segundos con saltos de 100 ms de la siguiente manera:

- La primera hipótesis asume que las características de estas dos ventanas vienen de dos locutores distintos y por consiguiente, son representados por modelos diferentes.
- La segunda hipótesis asume que las características de ambas ventanas pertenecen a un mismo hablante y por consiguiente, se pueden representar por el mismo modelo.

Como resultado de la comparación de estas dos hipótesis se tiene una curva de distancia para cada posición de la ventana deslizante. Los máximos de dicha curva que sobrepasen un determinado umbral indicarán los puntos de cambios de locutor. En nuestro sistema, las dos técnicas empleadas son BIC y GLR.

Las características de cada una de las mitades de la ventana se modelan a través de una distribución normal multivariada (gaussiana) y mediante matriz de covarianza completa o matriz de covarianza diagonal, dependiendo de la configuración escogida. A continuación, para determinar si un máximo local de la curva corresponde a un cambio de hablante se hace la diferencia entre su valor y los valores mínimos contiguos. Si esa diferencia es mayor al producto de la desviación típica de la medida -σ- por un umbral prefijado -α- entonces el máximo indica un punto de cambio de hablante. En nuestro sistema de diarización, el umbral α=0.5.

Primeramente, para cada punto detectado de cambio, se obtiene la medida ΔBIC, que calcula la distancia entre dos conjuntos de características para determinar si se modelan mejor con una o dos distribuciones gaussianas. A continuación, se calcula para cada segmento, la medida de distancia para cada par de segmentos contiguos. Si la medida $\Delta BIC < 0$ dado un umbral de decisión λ, el punto de cambio se desecha y ambos segmentos se unen en uno sólo.

Esto se repite para el nuevo segmento y el posterior hasta que la distancia entre cada par de segmentos ha sido calculada.

3.2.4 Agrupamiento sobre i-vectors

La última etapa, y tras la segmentación, es el proceso de agrupamiento. En esta fase se asigna a cada segmento una etiqueta de identidad para determinar cuáles corresponden al mismo locutor y cuáles a diferentes locutores.

En nuestro sistema, este agrupamiento se fundamenta en la comparación de los i-vectors que conforman cada uno de los segmentos esperando que se obtengan mejores resultados que los anteriores.

Primeramente, para cada segmento homogéneo, proveniente de la salida de la etapa de segmentación se obtiene un i-vector que describe este segmento. Esto se consigue a través de la matriz de transformación T que da lugar a i-vectors que resumen la información acústica del segmento en un vector de dimensión reducida, concretamente 600x1.

Más detalladamente, el extractor de i-vectors, que es un bloque de procesado fundamentado en "análisis factorial", nos permite representar un conjunto de vectores de características en un i-vector que recoge la información discriminativa de locutor y canal de manera muy eficiente. Está compuesto por UBM, que es un modelo universal que representa la distribución general de las características de los locutores, y la matriz de transformación T.

Esto lo que origina los i-vectors definitivos a los que realizamos el proceso de whitening descrito en el Apartado 2.6.1.

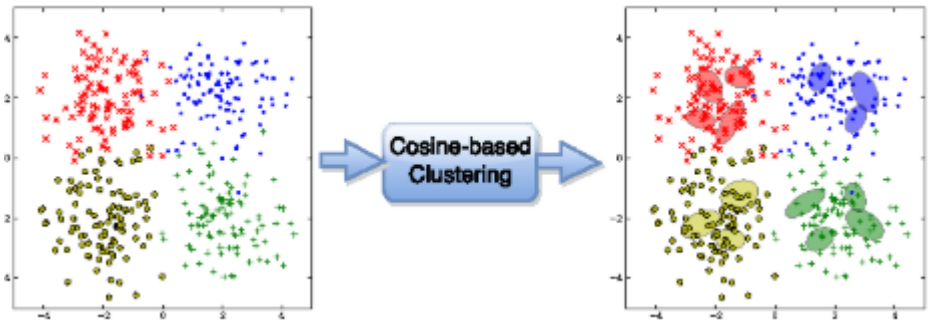

Figura 3-2: Agrupamiento basado en la medida de distancia coseno. | Fuente: [24]

Finalmente, se comparan los i-vectors procedentes de distintos segmentos a través de la distancia coseno. Para ello, en primer lugar se considera cada segmento de un cluster diferente y posteriormente se combinan siguiendo así la estrategia *bottom*-up, en función de una medida de distancia. Si esta medida es mayor que un determinado umbral, entonces los dos i-vectors pertenecen al mismo conjunto tal y como aparece en la **Figura 3-2**. Este es un proceso iterativo; en cada iteración se unen los dos clusters más cercanos y se evalúa el criterio de parada *cutoff*, y así sucesivamente mientras el criterio no se cumpla. Las técnicas que se utilizan en este Trabajo Fin de Grado para el agrupamiento son las detalladas en el Apartado 2.5.

4 Desarrollo

El desarrollo de este Trabajo Fin de Grado se puede distribuir en dos fases. Primeramente, se optimiza la etapa de clustering para mejorar su rendimiento considerando que el resto del sistema funciona 'perfecto', es decir, utilizando para el resto de módulos las etiquetas que contienen la información real para esas etapas (*ground truth*). Para ello, se realiza un barrido amplio de los parámetros de entrada del sistema de manera que se abarque la mayor cantidad de las distintas combinaciones de parámetros posibles.

A continuación, se evalúa el sistema de diarización de locutor por completo, es decir, el sistema funcionando en todas las etapas de tal manera que usando una configuración ya optimizada de las demás fases, la medida de error DER sea la mínima.

En este capítulo, se describe el proceso realizado en cada una de las fases anteriormente mencionadas, así como el entorno donde se han llevado a cabo los experimentos. Esto incluye las herramientas empleadas en el desarrollo del trabajo, la base de datos utilizada, tanto para el entrenamiento como para el desarrollo y evaluación, y las medidas de error que evalúan el sistema.

4.1 Entorno experimental

4.1.1 Equipo y programas informáticos

Los experimentos han sido realizados con un ordenador Pentium Dual-Core E6700 @ 3.10Ghz x2. Más concretamente con el sistema operativo Ubuntu en un servidor específico con dos microprocesadores Xeon E5649 a 2.53GHz y proporcionando 12 hilos de ejecución.

En el sistema de diarización de locutor se han empleado varias herramientas. Para la primera etapa de extracción de características se ha usado Kaldi. Esta es una herramienta bajo la licencia de Apache y es empleada para el reconocimiento de voz, siendo una de las que proporcionan un software más flexible y extensible [9].

En el resto del sistema, tanto para el entrenamiento de modelos como para las etapas que lo componen descritas en el Capítulo 2, se han empleado diferentes scripts realizados en MATLAB. Esto nos permite una implementación de algoritmos y una manipulación y representación de datos bastante sencilla.

En relación a la evaluación del rendimiento del sistema, se ha empleado un script proporcionado en la evaluación Albayzín. Dicho script ejecuta una herramienta externa (md-eval-v21) programada en Perl y que consiste en comparar las etiquetas reales (*ground truth*) con las etiquetas que nos proporciona la salida del sistema.

4.1.2 Base de datos

En nuestro sistema se han empleado dos bases de datos. La correspondiente con las noticias catalanas del canal de televisión 3/24 TV usada en la evaluación Albayzín 2010 y la donada por la Corporación Aragonesa de Radio y Televisión (CARTV) para la evaluación de diarización de locutor Albayzín 2016 [10]. Estos son datos para entrenamiento y desarrollo, y posteriormente, se entregan datos de evaluación sin etiquetar para los cuales hay que generar las etiquetas indicando los segmentos donde aparecen los locutores.

El audio proporcionado se suministra en formato PCM, monocanal, 16 bits de resolución y frecuencia de muestreo de 16000 Hz.

4.1.2.1 Entrenamiento

El primer conjunto de datos es la base de datos de noticias de difusión en catalán del canal de televisión 3/24 TV propuesto para la evaluación de segmentación de audio Albayzín 2010 [11]. Esta base de datos fue grabada por el Centro de Investigación TALP de la Universidad Politécnica de Cataluña.

La base de datos es empleada para entrenamiento -*train*- del extractor de i-vectors (UBM y matriz T), la normalización de whitening, la compensación LDA y WCCN, y las matrices de PLDA.

Esta base de datos contiene alrededor de 87 horas de grabaciones en las que la voz puede ser encontrada el 92% del tiempo. Con respecto a las clases superpuestas, la voz se puede encontrar junto con el ruido el 40% del tiempo y junto con la música el 15% como se puede apreciar en la **Figura 4-1**.

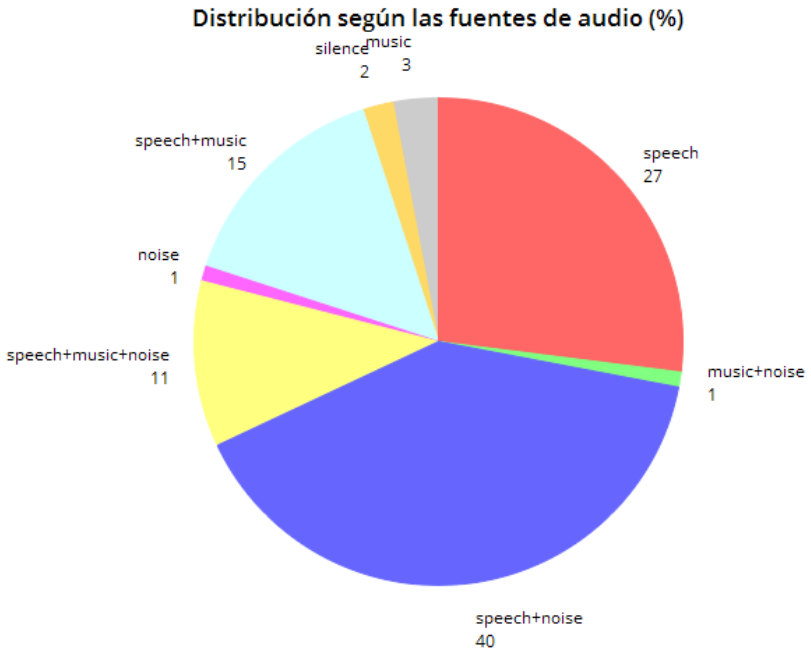

Figura 4-1: Distribución de la base de datos de 3/24 TV según su tipo de audio

4.1.2.2 Desarrollo y evaluación

Para esta base de datos, se distinguen datos de desarrollo *-devel-* y datos de evaluación final del sistema *-eval-*. Los primeros de ellos se usan para ajustar los parámetros de la etapa de clustering de manera que obtengamos el menor porcentaje de error de diarización. Y los datos restantes, sirven para evaluar el rendimiento del sistema ya con los mejores factores obtenidos tras los barridos paramétricos.

Este conjunto de datos contiene alrededor del 85% de voz, el 62% es música y 30% de ruido de tal forma que el 35% del audio contiene música junto con el habla, el 13% es ruido junto con la voz y el 22% es habla sólo [12]. Esta distribución según las fuentes de audio la podemos visualizar mejor en la **Figura 4-2**.

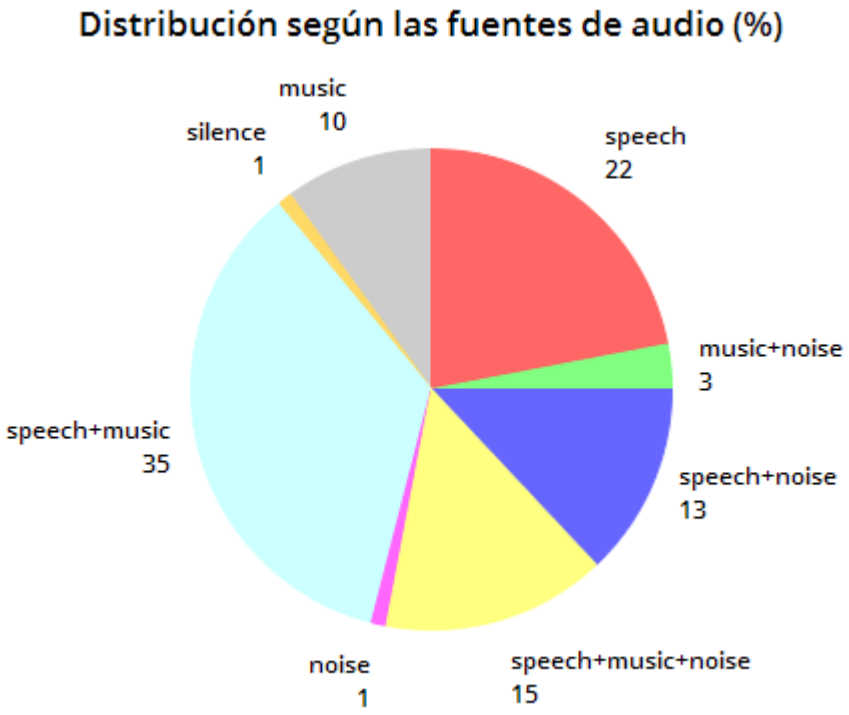

Figura 4-2: Distribución de la base de datos de CARTV según su tipo de audio

La base de datos cedida por CARTV tiene en su totalidad más de 20 horas de grabación en sus ficheros de audio, incluyendo ambas partes, es decir, tanto *-devel-* como *-eval-*. En la **Tabla 4-1** queda reflejado la información sobre locutores de la totalidad de la base de datos.

Base de datos	Nº de archivos	Duración	Nº medio de locutores por grabación	Nº medio de segmentos por grabación
Development	32	5h16m14s	7,9	44,1
Evaluation	72	17h58m48s	13,1	106,2

Tabla 4-1: Distribución cuantitativa de la base de datos

Se puede observar como en la partición de los datos, los relacionados con la evaluación son de mucho mayor volumen con respecto a los de desarrollo. Se requiere que nuestro sistema final tenga un elevado rendimiento y es necesario un análisis extenso de todas sus características en las distintas etapas con todos los ficheros de audio.

4.1.3 Métrica de evaluación

Para comprobar el rendimiento del sistema de diarización de locutor propuesto, el NIST[12] ha desarrollado una herramienta donde el funcionamiento se evalúa a través de una medida llamada tasa de error de diarización o DER.

La tasa de error de diarización se computará exactamente como la fracción de tiempo del hablante que no se atribuye correctamente a ese hablante específico, es decir, el tiempo total erróneamente etiquetado dividido por la duración total del audio hablado [13]. Esto se realizará para cada uno de los ficheros de la base de datos, incluyendo regiones donde hay más de un locutor presente. Por consiguiente, los segmentos de mayor duración tienen más influencia en el cálculo de la DER con respecto a los más cortos.

$$DER = \frac{speaker\ error\ +\ miss\ +\ false\ alarm}{total\ reference\ speech\ time} \tag{4.1}$$

Como podemos visualizar, la tasa de error de diarización depende a su vez de tres distintos tipos de error que se dan al etiquetar de manera incorrecta el audio.

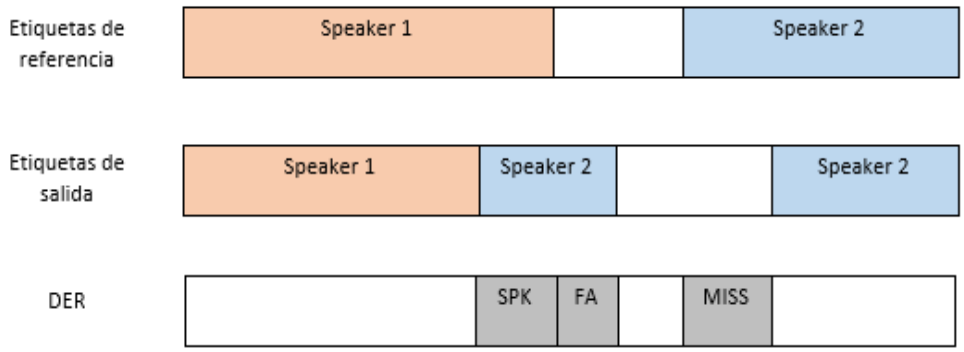

Figura 4-3: Cálculo de la medida *Diarization Error Rate* a partir de sus componentes

Estos tres errores se definen de la siguiente forma:

Speaker Error (SPK). Se corresponde con la cantidad de tiempo que se asigna a un locutor incorrecto. Este error suele ocurrir en segmentos donde el número de hablantes es mayor que el número de hablantes en las etiquetas de referencia, aunque también sucede si el número de hablantes es menor con respecto a la referencia. Este es el error más importante de los tres que componen la DER.

12 National Institute of Standards and Technology

False alarm (FA). Se define como el tiempo que un locutor ha sido etiquetado por el sistema de diarización pero en realidad no está presente en el segmento. Es decir, es la cantidad de tiempo detectado como habla cuando no lo es. Por ello, este error depende casi en su totalidad al detector de actividad.

Missed Speech (MISS). Es la componente que se refiere a la cantidad de tiempo que la voz está presente pero no está etiquetada por el sistema de diarización. Se puede entender como el inverso que el *FA* y al igual que este, depende en casi exclusivamente del VAD.

En la **Figura 4-4** se puede ver un ejemplo de la evaluación del sistema con esta métrica y la aparición de estas componentes.

```
*** Performance analysis for Speaker Diarization for ALL ***

    EVAL TIME =  18640.67 secs
  EVAL SPEECH =  16587.18 secs ( 89.0 percent of evaluated time)
  SCORED TIME =  18640.67 secs (100.0 percent of evaluated time)
SCORED SPEECH =  16587.18 secs ( 89.0 percent of scored time)
   EVAL WORDS =      0
 SCORED WORDS =      0         (100.0 percent of evaluated words)
---------------------------------------------
MISSED SPEECH =    415.93 secs (  2.2 percent of scored time)
FALARM SPEECH =   1100.37 secs (  5.9 percent of scored time)
 MISSED WORDS =      0         (100.0 percent of scored words)
---------------------------------------------
SCORED SPEAKER TIME =  16793.81 secs (101.2 percent of scored
speech)
MISSED SPEAKER TIME =    622.56 secs (  3.7 percent of scored
speaker time)
FALARM SPEAKER TIME =   1100.37 secs (  6.6 percent of scored
speaker time)
 SPEAKER ERROR TIME =   7395.26 secs ( 44.0 percent of scored
speaker time)
SPEAKER ERROR WORDS =      0         (100.0 percent of scored
speaker words)
---------------------------------------------
 OVERALL SPEAKER DIARIZATION ERROR = 54.29 percent of scored speaker
time  `(AL
---------------------------------------------
 Speaker type confusion matrix -- speaker weighted
  REF\SYS (count)      unknown               MISS
unknown                 193 /  76.6%         59 /  23.4%
  FALSE ALARM           497 / 197.2%
---------------------------------------------
 Speaker type confusion matrix -- time weighted
  REF\SYS (seconds)    unknown               MISS
unknown             16171.25 /  96.3%     622.56 /   3.7%
  FALSE ALARM        1100.37 /   6.6%
---------------------------------------------
```

Figura 4-4: Salida en el terminal de la evaluación del sistema

Se observa que el tiempo de voz es de 101,2%. Esto se debe a que hay solapamiento entre locutores por lo que evaluamos más cantidad de tiempo de la que contiene la base de datos (o fichero si es una prueba unitaria).

4.2 Evolución y particularidades de los experimentos

La primera etapa se centra en optimizar el módulo de clustering de i-vectors considerando que lo anterior funciona de manera 'perfecta' para el sistema de diarización de locutor.

Para ello, se realizan dos posibles casos en las etapas de detección de actividad y segmentación. En una de ellas se considera que el rendimiento de las etapas anteriores es perfecto, para evaluar el clustering de forma aislada; en la otra, se evalúa en condiciones reales, con una "entrada imperfecta" que es fruto de los errores existentes en etapas anteriores:

- *perfect.* Se cogen las etiquetas de referencia en dichas etapas, es decir, se trabaja con las *ground_truth* de modo que no actúen las funciones del detector de actividad y segmentación.
- *system.* Se leen unas etiquetas generadas con un sistema VAD, que puede tener errores, y se usan las funciones de segmentación, que tampoco son perfectas.

Esto nos permite aislar y tratar de manera casi independiente el agrupamiento de i-vectors del resto del sistema ya que a la salida de la segmentación, cuando se usa la opción *perfect* hay un único locutor por segmento, que es lo que queremos a la hora de evaluar el agrupamiento. Por consiguiente, si esta opción está activa la tasa de error de diarización por parte de estas dos etapas va a ser nulo (*DER = 0%*) ya que solo se estará procesando segmentos con voz y eso evitará ciertos errores (falsas alarmas y falsos rechazos) con lo que se puede optimizar el clustering de los i-vectors variando los parámetros que conforman dicha etapa.

4.2.1 Análisis de parámetros y scoring para agrupamiento de los i-vectors

Tras considerar que nuestro sistema funciona sin ningún tipo de error al realizar el caso *perfect* en las etapas previas al agrupamiento, se va a acometer una variación de los parámetros involucrados en el clustering de i-vectors que da a lugar a múltiples combinaciones que evaluaremos para dar con la configuración óptima con la base de datos de desarrollo -*devel*-:

- *cutoff.* Es un criterio o umbral empleado en la etapa de agrupamiento mediante el cual se decide si dos i-vectors pertenecen al mismo cluster. Es decir, es un parámetro que indica parada si los nuevos clusters distan de los clusters obtenidos en la iteración previa menos que una determinada distancia.
- *distance.* Es una medida de distancia (valor) que se emplea para cuantificar la diferencia entre segmentos. Hay múltiples opciones como 'euclidean', 'chebychev', 'correlation',... Se emplea la distancia 'cosine' porque lo que mejor determina la similitud entre locutores representados mediante i-vectors es el ángulo entre estos vectores. Si el coseno entre dos segmentos es similar a uno, indica que pertenecerán al mismo cluster.
- *linkage.* Es la técnica (método) empleada a la hora de comparar la distancia que hay entre dos cluster. Al igual que sucedía con el parámetro anterior, tiene varias opciones tales como:

- ‘single’. Considera la distancia entre dos cluster como la distancia entre los i-vectors más cercanos.
- ‘centroid’. Se basa en la distancia centroide.
- ‘median’. Se fundamenta en la distancia al centro de masas de un cluster.
- ‘average’. La elegida en nuestro trabajo que equivale a calcular las distancia entre cada par de i-vectors de uno y otro cluster, y hacer el promedio.

Estos son los parámetros generales que se pueden modificar en la etapa de clustering. Pero, asimismo, también hay un conjunto de técnicas y demás factores que dependen del tipo de scoring que se vaya a utilizar.

Particularizando en las técnicas, para los tres tipos de métodos de scoring vistos se ha hecho uso de whitening+lnorm. Esto se debe a que al aplicar ambos procedimientos, los i-vectors se distribuyen de mejor manera en el espacio dando lugar a tasas de error de diarización más bajas.

4.2.1.1 Otras casuísticas para LDA

Además de los parámetros descritos, que son comunes en los tres tipos de scoring, en esta técnica se puede probar con matrices LDA de distintas dimensiones. En nuestro sistema, inicialmente tenemos una matriz *A* de proyección que reduce la dimensión de los i-vectors a 100, pero podría funcionar con otro número de dimensiones. Es por ello, que se ha variado este factor para tratar de encontrar la mejor dimensionalidad para los vectores.

De igual manera, y tras desplazar los i-vectors al nuevo espacio, se emplea la compensación WCCN. Esta técnica se utiliza para compensar de forma distinta en cada dimensión la variabilidad entre distintos locutores. El número de dimensiones de la matriz WCCN está determinado por el número de dimensiones de los i-vectors; si se usa antes LDA, estará determinado por el número de dimensiones de la matriz LDA.

4.2.1.2 Otras casuísticas para PLDA

Como sucede con LDA, para la técnica de scoring PLDA también hay unos parámetros de entrada (a parte de los generales) modificables para determinar el ajuste más óptimo.

Estos parámetros son las matrices que modelan la variabilidad intra-locutor e inter-locutor, que inicialmente están a valor 100, pero pueden ser modificadas para encontrar la menor DER. Se emplean en la técnica de scoring GPLDA que devuelve scores, y se pasa a distancias sabiendo que distancia=-scores (se puede decir que es “lo contrario”) que es en base a lo que se hace el agrupamiento.

Para finalizar esta primera etapa, se ha probado la mejor configuración para cada técnica de scoring con los datos de *-devel-*, en la base de datos de *–eval-* para ver su comportamiento en un entorno de evaluación.

La segunda etapa consiste en probar el sistema funcionando en todas las fases, es decir, el sistema actuando en "condiciones normales". Para ello, se elige la opción '*system*' en los módulos de todo el sistema. Primero con la base de datos de *-devel-*, para buscar una configuración de la segmentación que dé lugar a los mejores resultados (como el VAD es un sistema independiente, no se ha configurado nada). Y con la mejor configuración obtenida para los datos de desarrollo, se prueba en *-eval-* tal y como se ha realizado con la etapa de clustering, con el objetivo de conseguir una tasa de error de diarización lo más baja posible.

4.2.2 Análisis de parámetros para la segmentación

Una vez observado el comportamiento del clustering de i-vectors como módulo independiente, se va a proceder a evaluar el sistema completo. Para esto, es necesario analizar los parámetros de las etapas que anteriormente se asumían que eran 'perfectas' como es la segmentación. Los factores que se han modificado son:

- *Tipo de técnica empleada.* Señala el tipo de procedimiento utilizado para las hipótesis sobre parejas de ventanas deslizantes decidiendo si las características de las dos ventanas proceden de dos locutores diferentes y por tanto, son representados por distintos modelos, o las características de las dos ventanas proceden de un mismo locutor, en cuyo caso se representa por el mismo modelo. Se distinguen dos tipos:
  - Δ*BIC.* Es una medida que calcula la distancia entre dos conjuntos de características para determinar si se modelan mejor con una misma distribución gaussiana o con dos distintas [14].
  - *GLR.* Esta técnica es un ratio de verosimilitudes o cociente entre la probabilidad de un conjunto de características a ser modelado por una sola distribución gaussiana al provenir de un mismo locutor y, la probabilidad de un conjunto de características a ser modelado por diferentes gaussianas al considerarse distintos locutores [15].
- *window.* Este parámetro indica el tamaño de la ventana que se emplea para recorrer la señal de audio.
- *lambda (λ).* Este parámetro indica el umbral de decisión para la técnica Δ*BIC.*
- *Tipo de matriz de covarianza.* Este parámetro se emplea para el modelado de las características de cada ventana pudiendo ser de dos tipos:
  - *full.* En esta opción se emplea la matriz de covarianza gaussiana completa.
  - *diag.* En esta opción se utiliza la diagonal principal de la matriz de covarianza gaussiana.

Con estas distintas combinaciones de parámetros, tenemos múltiples ejecuciones de nuestro sistema que da lugar a DER de muy distinto valor como veremos en el Capítulo 5. En la **Figura 4-5**, se tiene una representación acerca del trabajo realizado en esta segunda etapa.

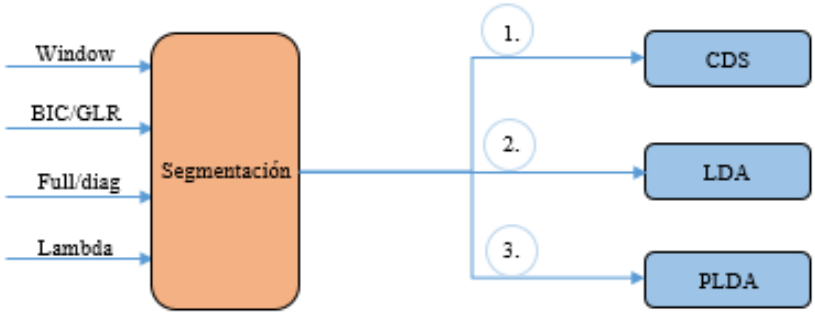

Figura 4-5: Representación de la segunda etapa de estudio del sistema

Se puede observar cómo una vez evaluado todos los parámetros de entrada a la etapa de segmentación, se procede a analizar dichas configuraciones con el sistema completo para cada técnica de scoring. Con esto tenemos el comportamiento real de nuestro sistema en una situación normal para los datos de evaluación.

5 Integración, pruebas y resultados

En este capítulo se incluyen los resultados acerca de los experimentos y las pruebas realizadas que tienen como objetivo disminuir la tasa de error de diarización de nuestro sistema.

5.1 Evaluación de las técnicas de scoring

Como se indicó en el Capítulo 4, primeramente se focaliza en la etapa de agrupamiento de i-vectors con las demás funcionalidades trabajando de manera 'perfecta'. Es decir, utilizando las etiquetas de referencia para medir el rendimiento exclusivamente de la etapa de clustering y conocer el impacto de los parámetros involucrados sobre la DER.

5.1.1 CDS

En primer lugar, se va a modificar el cutoff a la vez que evaluamos el efecto de la normalización. Esto es, emplear whitening y lnorm. Debido a su gran interés e importancia, se va a ver el efecto de cada una de ellas por separado, inicialmente para los datos de -*devel*-, siendo estos los resultados obtenidos:

Cutoff	Sin Whitening y L-norm	Sólo Whitening	Sólo L-norm	L-norm+Whitening	Whitening+L-norm
0.25	51.16%	49.60%	51.16%	51.25%	49.60%
0.5	51.16%	49.60%	51.16%	51.25%	49.60%
0.75	38.90%	34.11%	38.90%	37.82%	34.11%
1.1	31.85%	29.78%	31.85%	31.13%	29.78%
1.5	59.30%	59.30%	59.30%	59.30%	59.30%

Tabla 5-1: Tasa de error de diarización DER en función de la normalización para los datos de desarrollo

Para los mínimos hallados en los dos casos, se realiza un barrido de cutoff "más fino" de manera que se cubra la opción de poder encontrar un valor más bajo.

Cutoff	Whitening+L-norm	Sin Whitening y L-norm
0.8	32.67%	38.47%
0.9	31.69%	38.91%
1.0	30.84%	37.74%
1.2	59.30%	59.30%
1.3	59.30%	59.30%

Tabla 5-2: Tasa de error de diarización DER para valores de cutoff adyacentes a los mínimos encontrados para los datos de desarrollo

A la vista de los resultados, el parámetro que mayor afecta a la tasa de error de diarización es el *cutoff*. Para valores situados en los extremos, se obtiene un peor rendimiento del sistema debido a que seríamos muy tolerantes (valores bajos) o muy restrictivos (valores altos) a la hora de decidir si dos i-vectors pertenecen al mismo cluster.

Con respecto a la normalización, como era de esperar, vemos que aplicar whitening+lnorm (en este orden) se obtienen mejores resultados, concretamente 29,78% frente a un 31.85% sin normalización para el mejor caso. Esto se debe ya que al emplear estas dos técnicas, los i-vectors son distribuidos de una mejor manera en espacios de alta dimensionalidad como es nuestro caso.

Teniendo en cuenta el análisis realizado con los datos de desarrollo, se va a probar la mejor configuración para los datos de -*eval*-.

Cutoff	Whitening+L-norm
1.1	40.94%

Tabla 5-3: Tasa de error de diarización DER para la mejor configuración con los datos de evaluación (CDS)

Si comparamos con el resultado anterior vemos que hay un claro aumento llegando hasta 40.94%, lo que significa un 11.16% de diferencia. Esto es debido a que los ficheros de desarrollo -*devel*- tienen duraciones muy distintas a las de evaluación -*eval*-, además de distinto número de locutores, como vemos en la **Tabla 4-1**.

5.1.2 LDA

LDA es una técnica de reducción de dimensionalidad. También se dice que es una técnica de "compensación de variabilidad" en el sentido que retiene sólo las dimensiones que varían de un locutor a otro. Por ello, se va a trabajar modificando las matrices *A* de proyección y *B* de compensación (LDA+WCCN) sabiendo del caso anterior que la aplicación de whitening+lnorm es la que mejores resultados produce, por tanto, partimos de esta premisa para todas las pruebas realizadas. Los resultados conseguidos aplicando en todos los casos ambas técnicas (LDA+WCCN) para distinto número de dimensiones y valores de *cutoff* son:

Cutoff \ Tamaño matriz A,B	50	100	150	200
0.25	50.51%	50.53%	50.27%	50.01%
0.5	50.51%	50.53%	50.27%	50.01%
0.75	39.57%	39.91%	40.12%	38.72%
1.1	32.50%	30.04%	32.77%	34.91%
1.5	59.30%	59.30%	59.30%	59.30%

Tabla 5-4: Tasa de error de diarización DER con distinto número de dimensiones del espacio proyectado para los datos de desarrollo (LDA)

Al igual que sucedía con CDS, la tasa de error de diarización más baja se da cuando el valor del *cutoff* es igual a 1,1. Concretamente es de 30,04% cuando se reduce la dimensionalidad a 100. Esta reducción permite que los i-vectors se ajusten al espacio de mejor manera.

Hay un punto *(lda_dim=250)* a partir del cual no hay variación de DER, siendo peor el rendimiento del sistema.

Una vez hecho el análisis paramétrico para la base de datos de desarrollo, se procede a ejecutar la base de datos de evaluación con los mejores valores obtenidos.

Tamaño matriz A,B / Cutoff	100
1.1	43.39%

Tabla 5-5: Tasa de error de diarización DER para la mejor configuración con los datos de evaluación (LDA)

El resultado obtenido es de 43.39% que se traduce en un 13.35% de empeoramiento. Del mismo modo que ocurre con CDS, al emplear los datos *-eval-* con esta técnica se logran resultados mediocres. Esto se debe a que las matrices *A* y *B* son entrenadas con los datos de *-train-* que son proporcionados en catalán mientras que la evaluación es con audio en castellano lo que complica y mucho el rendimiento del sistema al haber tanto desajuste por tratarse de diferentes idiomas que presentan una pronunciación distinta, entonación totalmente distinta, vocabulario diferente, etc.

5.1.3 PLDA

PLDA es una versión probabilística del anterior método. En esta técnica, primero se entrena las matrices de intra-variabilidad e inter-variabilidad de locutor a través de la técnica de training GPLDA. Estos son los parámetros sobre los que se va a trabajar para optimizar el sistema en esta técnica.

Cabe destacar que para realizar el scoring se emplea el método de scoring GPLDA, que es una medida de similitud entre i-vectors y calcula los scores de todas las combinaciones entre pares de i-vectors.

Los resultados conseguidos con esta técnica son los que se indican a continuación (en todos los casos se ha realizado whitening+lnorm):

Cutoff \ Tamaño matrices	50	100	150	200
0.25	54.00%	53.04%	51.93%	52.86%
0.5	54.00%	53.04%	51.93%	52.86%
0.75	34.05%	36.80%	33.77%	34.04%
1.1	32.97%	32.83%	28.60%	28.86%
1.5	59.30%	59.30%	57.96%	59.30%

Cutoff \ Tamaño matrices	250	300	350	400
0.25	52.49%	52.49%	52.49%	52.49%
0.5	52.49%	52.49%	52.49%	52.49%
0.75	32.73%	32.06%	32.57%	32.44%
1.1	33.71%	34.57%	34.70%	34.44%
1.5	59.30%	59.30%	59.30%	59.30%

Tabla 5-6: Tasa de error de diarización DER con distintos tamaños de matrices para los datos de desarrollo (PLDA)

Comparando los resultados de esta técnica con las demás, obtenemos el mínimo general de 28,60% lo que supone una diferencia de mejoría de 1,18% y 1,44% con respecto a CDS y LDA. Esta técnica es más compleja que las anteriores y mejora los resultados al emplear modelos probabilísticos para las características de los i-vectors.

En torno al valor 250 de tamaño de matriz en adelante, los mínimos locales no se encuentran cuando el cutoff es de 1,1 sino que están cuando el valor es de 0,75.

A continuación, se realizó una prueba en el entrenamiento de las matrices, ajustando el número de dimensiones al número de segmentos que llegan a la etapa de clustering. Para ello se modificó el script de entrenamiento consiguiendo los siguientes resultados:

Cutoff	Antes	Ahora
1.1	28.60%	47.99%

Tabla 5-7: Tasa de error de diarización DER con tamaño de matriz adaptativo para los datos de desarrollo

El resultado obtenido empeora el rendimiento debido a que mucha de las grabaciones tiene un número de segmentos bajos por lo que reducir tanto la dimensión provoca discordancia en los i-vectors.

Con los parámetros que mejor DER hemos tenido, se vuelcan a la base de datos de evaluación obteniendo los datos siguientes:

Cutoff \ Tamaño matrices	150
1.1	55.63%

Tabla 5-8: Tasa de error de diarización DER para la mejor configuración con los datos de evaluación (PLDA)

En este caso, que el entrenamiento se realice con audio en otro idioma con respecto a la evaluación conlleva un resultado bastante peor (27,03% de diferencia) puesto que las matrices de intra e inter-variabilidad del modelo resultan de este entrenamiento.

5.2 Evaluación del sistema de diarización de locutor completo configurando la segmentación

En los siguientes apartados, se muestran los datos correspondientes a la segunda etapa del Trabajo Fin de Grado donde se exponen las modificaciones realizadas en los parámetros dependientes de la segmentación para analizar la DER a la salida de cada uno de los tres casos de agrupamiento de i-vectors vistos.

Para todos los casos, el tipo de técnica empleada es ΔBIC con la matriz de covarianza diagonal, que como vemos en [19] da lugar a los mejores resultados.

5.2.1 CDS

Como estudiamos en el Apartado 5.1.1, la mejor configuración para esta técnica es cuando el valor de *cutoff* es de 1,1 y aplicando whitening+lnorm. Fijando dicha configuración, en las siguientes tablas se indican los datos obtenidos tras el barrido realizado de los parámetros de entrada de segmentación (tamaño de ventana, lambda):

Window \ λ	0.5	0.6	0.7	0.8
3	44.67%	44.07%	44.93%	46.82%
3.5	45.14%	44.79%	45.35%	44.75%
4	48.47%	46.57%	50.17%	48.63%
4.5	46.30%	45.25%	47.64%	51.47%

Tabla 5-9: Tasa de error de diarización DER configurando la segmentación con los datos de desarrollo (CDS)

En este caso, el mejor resultado se obtiene con un tamaño de ventana de 3 segundos y lambda = 0.6. Pero en general, podemos observar como las diferencias son mínimas entre distintas configuraciones (7,4% entre máximo y mínimo) al contrario que el resultado del apartado anterior (29,52% de desigualdad). Esto indica la mayor dependencia de los

parámetros de agrupamiento, es decir, la etapa de clustering de i-vectors es la que mayor impacto tiene sobre la DER.

Las imperfecciones de las etapas de detección de actividad y segmentación, que introducen errores, provocan el aumento de la tasa de error de diarización. Con la misma configuración, se obtiene un 44.07%, lo que supone una diferencia negativa de 14,29% con respecto a la evaluación aislada de CDS, que se traduce en un rendimiento claramente inferior.

Si trasladamos esta configuración sobre la base de datos de *-eval-*, alcanzamos los siguientes resultados:

Window \ λ	0.6
3	54.35%

Tabla 5-10: Tasa de error de diarización DER sobre el sistema de diarización completo con los datos de evaluación (CDS)

Por último, como era de esperar, el error introducido en las etapas que antes se configuraban con las etiquetas reales, provoca un aumento considerable en la DER (54,45% frente a 40,94%).

5.2.2 LDA

En esta técnica, al igual que en la anterior, se tienen unos parámetros fijos (*cutoff*, aplicación de whitening+lnorm) de la etapa de agrupamiento de i-vectors como resultado del análisis realizado anteriormente. Únicamente, se añade las matrices *A* y *B* a dicha configuración. Modificando las variables de segmentación conseguimos los siguientes datos:

Window \ λ	0.5	0.6	0.7	0.8
3	47.99%	46.20%	46.58%	45.73%
3.5	46.09%	45.48%	45.61%	46.51%
4	45.22%	48.58%	45.52%	47.64%
4.5	49.32%	45.90%	49.21%	48.10%

Tabla 5-11: Tasa de error de diarización DER configurando la segmentación con los datos de desarrollo (LDA)

En la tabla anterior se puede comprobar como los resultados son aún más próximos entre sí. La variación entre el rango seleccionado de ventana y lambda apenas influye en la DER (4.1% de diferencia entre extremos), siendo el *cutoff*, de nuevo, el parámetro más decisivo.

Para la base de datos de evaluación obtenemos los datos adjuntos en la tabla posterior:

Window \ λ	0.5
4	53.58%

Tabla 5-12: Tasa de error de diarización DER sobre el sistema de diarización completo con los datos de evaluación (LDA)

LDA es una técnica cuyo objetivo es que los resultados de cosine distance sean mejores. Comparando con la **Tabla 5-10**, se produce una leve mejoría aunque la tasa sigue siendo alta al entrenar al extractor de i-vectors con grabaciones en otro lenguaje.

5.2.3 PLDA

PLDA se basa en modelos probabilísticos, donde los factores intra-variabilidad e inter-variabilidad de locutor son los principales a la hora de conformar dicha técnica. Si unificamos estos parámetros con los de segmentación, conseguimos los siguientes resultados indicados en la tabla:

Window \ λ	0.5	0.6	0.7	0.8
3	56.62%	55.68%	55.60%	56.49%
3.5	53.67%	52.67%	52.32%	51.43%
4	56.89%	52.89%	54.70%	55.92%
4.5	54.33%	53.71%	53.31%	54.29%

Tabla 5-13: Tasa de error de diarización DER configurando la segmentación con los datos de desarrollo (PLDA)

En los datos, se puede concluir que con esta técnica obtenemos de media una tasa de DER mayor que para los demás casos, lo cual indica que la segmentación afecta en mayor medida a la entrada de la etapa de clustering.

Para los datos de -*eval*- se han realizado las siguientes mediciones con la configuración más óptima:

Window \ λ	0.8
3.5	63.23%

Tabla 5-14: Tasa de error de diarización DER sobre el sistema de diarización completo con los datos de evaluación (PLDA)

Las matrices de intra-variabilidad e inter-variabilidad de locutor se ven afectadas de mayor manera por los posibles fallos del detector de actividad y de la segmentación. Esto supone que sea la técnica que peor funcione en el sistema de diarización completo, con una tasa de error de 63,23%, muy superior al resto.

6 Conclusiones y trabajo futuro

6.1 Conclusiones

En este Trabajo Fin de Grado, se ha realizado el análisis de las técnicas de agrupamiento mediante i-vectors para los sistemas automáticos de seguimiento de locutor, concretamente el empleado en la evaluación Albayzín 2016.

Para el sistema dado, se ha optimizado la etapa de clustering realizando una investigación exhaustiva de los parámetros que componen dicha etapa y a su vez, de los parámetros que forman cada una de las técnicas vistas como son CDS, LDA y PLDA. En este contexto, el empleo de técnicas de normalización como son la aplicación de whitening a los i-vectors, normalización de longitud (lnorm) y compensación WCCN mejora el rendimiento del sistema consiguiendo tasas de error de diarización más bajas.

En referencia a las técnicas de clustering empleadas, el uso de PLDA es la que mejores resultados ha arrojado con los datos de -*devel*-, registrando el mayor rendimiento con una DER del 28,60%, pero es menos robusta que CDS al "desajuste", ya que los datos de desarrollo son distintos a los de evaluación (aunque sean ambos en castellano) en cuanto a duraciones, número de locutores, etc.

Una vez realizada la inspección del módulo de agrupamiento, se presenta un estudio del sistema de seguimiento de locutor funcionando en un escenario real. De acuerdo con esto, se han efectuado distintas configuraciones de la segmentación (el VAD al ser un módulo independiente, no se ha modificado). Este enfoque conlleva tasas más altas, puesto que las etapas no son perfectas e introducen errores. Más concretamente, la DER aumenta alrededor de un 15%. A pesar de ello, para diversos valores de los parámetros de segmentación, la variación de la DER es mínima lo que indica la fuerte dependencia de la etapa de agrupación en la tasa de error de diarización.

El hecho de que el extractor de i-vectors (modelo UBM y matriz T) y las técnicas que requieren entrenamiento usen una base de datos de entrenamiento, dada en la evaluación Albayzín 2010 que contiene audio en catalán, muy distinta de la base de datos de desarrollo y evaluación que es en castellano, limita en gran medida el rendimiento del sistema.

6.2 Trabajo futuro

Para dar continuidad al trabajo expuesto, a continuación se mencionan algunas líneas de actuación a seguir:

- Se podría analizar el empleo de modelos que no hagan uso de los coeficientes MFCC, sino de parámetros de alto nivel, como son los fonemas, palabras o patrones de pronunciación que mejoren el rendimiento del sistema.

- Se podría comparar a nivel de evaluación el algoritmo jerárquico empleado en la etapa de agrupación con otros tipos de algoritmo buscando así el que mejor se adapte a nuestro sistema.

- Otra posibilidad de mejora podría darse en el entrenamiento del extractor de i-vectors, de modo que se empleen audios más parecidos a los de desarrollo y evaluación con motivo de solventar el desajuste ocasionado.

- A la vista de los resultados, sería interesante la elaboración de algún módulo o aplicar técnicas de ajuste a las matrices PLDA que mejore el rendimiento del sistema.

- Sería de utilidad, analizar a nivel de código el sistema de diarización de locutor, de manera que se pueda optimizar para reducir los tiempos de ejecución.

Referencias

[1] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, vol 19, pp. 788-798.

[2] Juan Luis Navarro Mesa, Alfonso Ortega, Antonio Teixeira, Eduardo Hernández Pérez, Pedro Quintana Morales, Antonio Ravelo García, Iván Guerra Moreno, Doroteo T. Toledano. Advances in Speech and Language Technologies for Iberian Languages. November 2014.

[3] Andrew O. Hatch, Sachin Kajarekar, Andreas Stolcke. Within-Class Covariance Normalization for SVM-based Speaker Recognition. January 2006.

[4] Theodoros Giannakopoulos, Sergios Petridis. Detection and clustering of musical audio parts using Fisher Semi-Discriminant Analysis. pp. 1290-1293. August 2012.

[5] A. Kanagasundaram, D. Dean, R. Vogt, M. McLaren, S. Sridharan, M. Mason. Weighted LDA techniques for i-vector based speaker verification. pp. 4781-4784.

[6] Patrick Kenny, G. Boulianne, P. Oullet, P. Dumouchel. Join Factor Analysis versus Eigenchannels in Speaker Recognition. May 2007.

[7] Jan Silovsky, Jan Prazak, Petr Cerva, Jindrich Zdansky, Jan Nouza. PLDA-based Clustering for Speaker Diarization of Broadcast Streams (2011). pp. 2909-2912.

[8] [en línea]. Disponible en: http://multimedia.icsi.berkeley.edu/speaker-diarization/

[9] KALDI, [en línea]. Disponible en: http://kaldi-asr.org/doc/

[10] Albayzin Evaluation 2016, [en línea]. Disponible en: https://iberspeech2016.inesc-id.pt/index.php/albayzin-evaluation/

[11] T Butko, CN Camprubí, H Schulz, in II Iberian SLTech. Albayzin-2010 audio segmentation evaluation: evaluation setup and results (FALA Vigo, 2010), pp. 305–308

[12] Alfonso Ortega, Ignacio Viñals, Antonio Miguel, Eduardo Lleida. The Albayzin 2016 Diarization Evaluation. June 2016.

[13] Galibert, Olivier. (2013). Methodologies for the evaluation of Speaker Diarization and Automatic Speech Recognition in the presence of overlapping speech.

[14] Wu, Chung-Hsien & Chiu, Yu & Shia, Chi-Jiun & Lin, Chun-Yu. (2006). Automatic segmentation and identification of mixed-language speech using Delta-BIC and LSA-Based GMMs. Audio, Speech, and Language Processing, IEEE Transactions on. 14. pp. 266 - 276.

[15] Wang, David, Vogt, Robert J., Mason, Michael W., & Sridharan, Sridha. Automatic Audio Segmentation Using the Generalized Likehood Ratio. 2008.

[16] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, S. Meignier. "An Open-source State-of-the-art Toolbox for Broadcast News Diarization," Interspeech, Lyon (France), 25-29 Aug. 2013

[17] Kyu J. Han, Samuel Kim, Shrikanth S. Narayanan. "Strategies to improve the robustness of Agglomerative Hierarchical Clustering under data source variation for Speaker Diarization". pp. 1590-1601. Nov. 2008.

[18] Naranjo, Benjamín García. "Segmentación de audio broadcast". Enero 2016.

[19] Suárez Pedrero, Guillermo. "Segmentación no supervisada de señales de audio y voz". Junio 2018.

[20] Jin X., Han J. (2011) "Expectation Maximization Clustering". In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA

[21] Campbell, W.M., Sturim, D.E., Reynolds, D.A., & Solomonoff, A. (2006). "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation". 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 1, I-I.

[22] Cariad, Chester and Holden T. Meacker. "Algorithmic Tools for Mining High-Dimensional Cytometry Data". pp. 773-779. Aug 2015.

[23] Mohammadi, Mahdi & Al-Azab, Fadwa & Raahemi, Bijan & Richards, Gregory & Jaworska, Natalia & Smith, Dylan & de la Salle, Sara & Blier, Pierre & Knott, Verner. (2015). Data mining EEG signals in depression for their diagnostic value. BMC medical informatics and decision making.

[24] Khoury, E., Shafey, L.E., & Ferras, M. (2014). "Hierarchical speaker clustering methods for the NIST i-vector Challenge".